Large Scale Model-Driven comparison of Metagenome Assembled Genomes from Diverse Environments

Stephen Nayfach[1], José P. Faria[2*](jplfaria@anl.gov), Simon Roux[1], Rekha Seshadri[1], Daniel Udwary[1], Neha Varghese[1], Frederik Schulz1, Dongying Wu[1], David Paez-Espino[1], I-Min Chen[1], Marcel Huntemann[1], Krishna Palaniappan[1], Joshua Ladau[1], Supratim Mukherjee[1], T.B.K. Reddy[1], Torben Nielsen[1], Edward Kirton[1], Janaka N. Edirisinghe[2], Christopher S. Henry[2], Sean P. Jungbluth[3], Dylan Chivian[3], Paramvir Dehal[3], Elisha M. Wood-Charlson[3], Adam P. Arkin[3], Susannah Tringe[1], Axel Visel[1], IMG/M Data Consortium, Tanja Woyke[1], Nigel J. Mouncey[1], Natalia N. Ivanova[1], Nikos C. Kyrpides[1], Emiley A. Eloe-Fadrosh[1]

[1]DOE Joint Genome Institute, Berkeley, California, USA

[2]Argonne National Laboratory, Argonne, Illinois, USA

[3]Lawrence Berkeley National Laboratory, Berkeley, California, USA

https://genome.jgi.doe.gov/portal/GEMs/GEMs.download.html

https://narrative.kbase.us/#org/jgimags

Project goals:

Over 10,000 metagenomes collected from diverse habitats covering all of Earth's continents and oceans, human- and animal-host associated microbiomes, engineered environments, and natural and agricultural soils were used to generate over 52,000 metagenome-assembled genomes (MAGs). Metabolic models were constructed using the high-quality non-redundant MAGs to explore the distribution of metabolic functions across ecosystems. To evaluate the quality of the constructed metabolic models, species-level genomic references were compared and resulted in a high correlation of predicted functions, indicating the MAGs encode representative pathways to their isolate counterparts.

The reconstruction of bacterial and archaeal genomes from shotgun metagenomes provides insight into the ecology and evolution of environmental and host-associated microbiomes. Genome-scale metabolic models were built and analyzed for the non-redundant, high quality GEMs with >40 representatives per environment (n = 3255) using the ModelSEED[1] pipeline in KBase[2] (See supplemental materials). In brief, GEMs from similar environments, such as human and mammal, were shown to cluster by pathway presence (containing at least one complete flux pathway) implying that pathways are differentially sorted by distinct environmental factors. For the 607 GEMs with close (>95% ANI) RefSeq genomes identified, a comparison of GEM to RefSeq models revealed a very high (>0.98) correlation suggesting that these high-quality

GEMS are very near complete and representative of their full metabolic potential. To demonstrate that the high correlation was not the result of all models being similar, correlations were also computed for random pairs of GEM models and RefSeq models, resulting in a much lower correlation of 0.83. All GEMs, associated RefSeq genomes, and metabolic models are freely available in the "JGI MAG Database" KBase organization (https://narrative.kbase.us/#org/jgimags). To validate the high-quality GEMs metabolic models, pathway presence profiles were computed for reference genomes associated with humans and the built environment, as these two environments have >100 GEMs with associated reference genomes. The resulting profiles were nearly identical for all pathways. Pearson correlation coefficients were calculated for each GEM and corresponding reference genome across 55 metabolic pathways, with an average value >0.98. When the GEM and reference genomes were randomly paired and a Pearson correlation was calculated, the average correlation dropped to ~0.82, indicating that the high correlation previously reflects the similarity of the GEM and reference genome.

References

1. Arkin, Adam P., Robert W. Cottingham, Christopher S. Henry, Nomi L. Harris, Rick L. Stevens, Sergei Maslov, Paramvir Dehal, et al. 2018. "KBase: The United States Department of Energy Systems Biology Knowledgebase." *Nature Biotechnology* 36 (7): 566–69.

2. Henry, Christopher S., Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens. 2010. "High-Throughput Generation, Optimization and Analysis of Genome-Scale Metabolic Models." *Nature Biotechnology* 28 (9): 977–82.

KBase is funded by the Genomic Science program within the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under award numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.